AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews

Shea, B.J.; Hamel, C.D.; Wells, G.A.; Bouter, L.M.; Kristjansson, E.; Grimshaw, J.; Henry, D.A.; Boers, M.

published in
Journal of Clinical Epidemiology
2009

DOI (link to publisher)
10.1016/j.jclinepi.2008.10.009

document version
Publisher's PDF, also known as Version of record

ELSEVIER

Journal of Clinical Epidemiology 62 (2009) 1013−1020

Journal of Clinical Epidemiology

AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews

Beverley J. Shea[a,b,c,*], Candyce Hamel[a], George A. Wells[d,e], Lex M. Bouter[b], Elizabeth Kristjansson[f], Jeremy Grimshaw[g], David A. Henry[h], Maarten Boers[c]

[a]*Community Information and Epidemiological Technologies (CIET), Institute of Population Health, Ottawa, Ontario, Canada*
[b]*Institute for Research in Extramural Medicine (EMGO Institute), VU University Medical Center, Amsterdam, The Netherlands*
[c]*Department of Clinical Epidemiology and Biostatistics, VU University Medical Center, Amsterdam, The Netherlands*
[d]*Ottawa Heart Institute, University of Ottawa, Ottawa, Ontario, Canada*
[e]*Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada*
[f]*Faculty of Social Sciences, School of Psychology, University of Ottawa, Ontario, Canada*
[g]*Clinical Epidemiology Program, Department of Medicine, Ottawa Health Research Institute, University of Ottawa, Ottawa, Ontario, Canada*
[h]*Institute for Clinical Evaluative Sciences, Toronto, Ontario, Canada*

Accepted 29 October 2008

Abstract

Objective: Our purpose was to measure the agreement, reliability, construct validity, and feasibility of a measurement tool to assess systematic reviews (AMSTAR).

Study Design and Setting: We randomly selected 30 systematic reviews from a database. Each was assessed by two reviewers using: (1) the enhanced quality assessment questionnaire (Overview of Quality Assessment Questionnaire [OQAQ]); (2) Sacks' instrument; and (3) our newly developed measurement tool (AMSTAR). We report on reliability (interobserver kappas of the 11 AMSTAR items), intraclass correlation coefficients (ICCs) of the sum scores, construct validity (ICCs of the sum scores of AMSTAR compared with those of other instruments), and completion times.

Results: The interrater agreement of the individual items of AMSTAR was substantial with a mean kappa of 0.70 (95% confidence interval [CI]: 0.57, 0.83) (range: 0.38−1.0). Kappas recorded for the other instruments were 0.63 (95% CI: 0.38, 0.78) for enhanced OQAQ and 0.40 (95% CI: 0.29, 0.50) for the Sacks' instrument. The ICC of the total score for AMSTAR was 0.84 (95% CI: 0.65, 0.92) compared with 0.91 (95% CI: 0.82, 0.96) for OQAQ and 0.86 (95% CI: 0.71, 0.94) for the Sacks' instrument. AMSTAR proved easy to apply, each review taking about 15 minutes to complete.

Conclusions: AMSTAR has good agreement, reliability, construct validity, and feasibility. These findings need confirmation by a broader range of assessors and a more diverse range of reviews.

Keywords: Systematic reviews; Meta-analysis; Methodological quality; Validity; Reliability; Feasibility

1. Background

Systematic reviews have become the standard approach in assessing and summarizing applied health research, but the quality of systematic reviews has received relatively little attention. Quality can be defined as the likelihood that the design of a systematic review will generate unbiased results [1].

Systematic reviews have appeared in medical journals since the late 1970s. Thousands of systematic reviews are available on all areas of health care, and a substantial portion of them has been produced by the Cochrane Collaboration. High methodoliogical quality is a pre-requisite for valid interpretation and application of review findings. However, systematic reviews are complex exercises, and assessing quality can be a daunting task. Clinicians and policy makers require guidance, which is not provided adequately by the available literature on the quality of systematic reviews. In a previous study, we summarized this literature, tested quality assessment tools, and reached the conclusion that current instruments for conducting methodological quality assessments of systematic reviews were suboptimal and needed revision and updating [2]. No single instrument has achieved dominance in terms of general use.

* Corresponding author: CIET Institute of Population Health, 1 Stewart Street, Room 319, Ottawa, Ontario, K1N 6N5, Canada. Tel.: +613-562-5800 ext. 8571; fax: +613-562-5392.

E-mail address: bshea@ciet.org or bevshea@uottawa.ca (B.J. Shea).

0895-4356/09/$ − see front matter
doi: 10.1016/j.jclinepi.2008.10.009

What is new?

- AMSTAR, a new instrument for evaluating systematic reviews, is reliable, valid, and easy to use.
- Currently, there is no agreement on which instrument to use when measuring the quality of systematic reviews. AMSTAR, a development of existing instruments, provides a possible solution.
- The instrument should be widely evaluated to confirm the performance metrics recorded here. The instrument should be updated as new knowledge is generated regarding factors that affect the quality of systematic reviews.

One popular instrument (QUality Of Reporting Of Meta-analyses [QUORUM]) is a reporting checklist, not a methodological quality assessment instrument.

Our review revealed a variety of weaknesses in the available instruments [2]. Our intention was not to come up with a truly novel approach, but to bring clarity to the field by: reviewing the available instruments, further developing and updating existing instruments, and providing a model that was validated and useable (in terms of comprehensibility and acceptable time for completion). Based on this evaluation, we created a measurement tool to assess systematic reviews (AMSTAR). This refines and enhances work presented in previously published instruments (by Oxman and Guyatt, 1991 [3] and Sacks et al., 1987 [4]) [5]. The present study concerns the internal validation of AMSTAR using the set of reviews used in its development. Here we focus on parameters of agreement, reliability; construct validity, and feasibility through comparisons with other instruments. An external validation of AMSTAR has been reported separately [6].

2. Methods

We used a computer-generated random sample of 30 (20%) of 151 systematic reviews that were used in the development of the instrument [5]. This sample contained 11 Cochrane and 19 non-Cochrane reviews, including meta-analyses and qualitative reviews. The topics of the reviews ranged across the spectrum of medicine [7–36]. Two reviewers (one without formal training) applied the new AMSTAR instrument and the two quality assessment tools, the enhanced Overview of Quality Assessment Questionnaire (OQAQ, originally developed by Oxman and Guyatt), and the instrument developed by Sacks et al. to all 30 reviews (C.H., B.J.S.) [3,4]. For each reviewer, the data set extracted contained three quality ratings for each review, yielding a total of six ratings per review.

2.1. Agreement and reliability

We calculated overall agreement and Cohen's kappa for each item (''yes'' scores vs. any other scores) [37]. Bland and Altman's limits of agreement method explained the agreement graphically [38–40]. We awarded each item scoring ''yes'' one point and summed these to calculate a total score. Intraclass correlation coefficients (ICCs) assessed the reliability of this total score [41]. We further scrutinized items and reviews with kappa values below 0.50. Finally, we repeated the exercise for the OQAQ and Sacks' instruments. Kappa values of less than 0 were rated as less than chance agreement; 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–0.99, almost perfect agreement. [42]

2.2. Construct validity

The new instrument already has high face and content validity by virtue of its construction process [5]. In the current study, we assessed construct validity by converting the mean total score (mean of two raters C.H. and B.J.S.) of each of the 30 reviews to a percentage of the maximum score for each of the three instruments. ICCs then assessed convergence of the total scores between each pair of instruments (AMSTAR–OQAQ, AMSTAR–Sacks, and OQAQ–Sacks).

2.3. Feasibility

Based on a guideline for assessing feasibility of instrument use developed by the Outcome Measures in Rheumatology (OMERACT) group [43], we compared the feasibility of the new instrument with that of the existing instruments by recording the time it took to complete scoring and the instances where scoring was difficult or impossible. The wording of individual items is critical for the performance of AMSTAR and fine-tuning is expected to be an ongoing task.

SPSS (version 13; SPSS Inc., Chicago, IL, USA) and MedCalc Software (Mariakerke, Belgium) were used to analyze the data, and the results were expressed as means and 95% confidence intervals (CIs) unless otherwise noted.

3. Results

The sample of 30 reviews adequately covered a wide range of quality, albeit with some underrepresentation of poor-quality reviews. Overall quality scores on AMSTAR ranged from 3 to 10 (out of a maximum of 11) with a flat distribution between 3.5 and 10 and a mean percentage score of 49.4%. The overall quality scores on Sacks' instrument ranged from 5 to 16 (out of a maximum score of 24), with a mean percentage score of 41.6%, and for OQAQ, scores ranged from 3 to 10 (out of a maximum score of 10), with a mean percentage score of 63.3%.

3.1. Agreement and reliability

The interobserver agreement of the individual items in the AMSTAR was high: mean = 0.88 (range: 0.73–1.0) with a mean kappa of 0.70 (95% CI: 0.57, 0.83) (range: 0.38–1.0). However, items 4 (publication status), 7 (report of assessment of scientific quality), and 9 (appropriate method to combine studies) scored fair to moderate at 0.38, 0.42, and 0.45, respectively. On the first two of these items, overall agreement was substantial at 0.80 and the relatively low kappa may be explained by a skewed distribution, that is, a high number of reviews in which the reviewers agreed on the score "no" (item 4) and "yes" (item 7), respectively. On item 8, overall agreement was also satisfactory at 0.74. Compared with the other instruments, agreement on individual items was similar to OQAQ: mean kappa of 0.63 (95% CI: 0.39, 0.78) (range 0.39–0.84), and superior to the Sacks' instrument: mean kappa of 0.40 (95% CI: 0.29, 0.50) (range: −0.47 to 0.93). In these instruments, fair to moderate agreement was also seen in the items covering assessment of scientific validity, statistical combinability, and comprehensive literature searching (Table 1).

For the AMSTAR total score, the mean difference between the two observers' scores was 0.2 (0.36–0.91). Agreement was similar in reviews with high- and low-quality scores (Fig. 1).

The interobserver ICC for the total score was excellent for all instruments: AMSTAR, 0.84 (95% CI: 0.65, 0.92); OQAQ, 0.91 (95% CI: 0.82, 0.96); and Sacks' instrument, 0.86 (95% CI: 0.71, 0.94). In one non-Cochrane review [12], observers differed by 3 points (6 vs. 9). In this review, the differences were noted on AMSTAR questions addressing duplication study selection and data extraction (item 2), publication status (item 4), and methods used to combine studies (item 9). In one Cochrane review [15], observers differed by 4 points (1 vs. 5). In this review, differences were noted on AMSTAR questions assessing the a priori design (item 1), publication status (item 4), scientific quality (item 7), and methods used to combine studies (item 9). The overall quality of Cochrane reviews included in this data set was somewhat higher than non-Cochrane reviews.

Table 1
Assessment of the interrater agreement for AMSTAR

Items	Kappa (95% CI)
1. Was an "a priori" design provided?	0.80 (0.63, 0.90)
2. Was there duplicate study selection and data extraction?	0.80 (0.17, 0.81)
3. Was a comprehensive literature search performed?	0.72 (0.40, 0.87)
4. Was the status of publication (i.e., grey literature) used as an inclusion criterion?	0.38 (0.28, 0.70)
5. Was a list of studies (included and excluded) provided?	0.56 (0.07, 0.79)
6. Were the characteristics of the included studies provided?	0.74 (0.45, 0.86)
7. Was the scientific quality of the included studies assessed and documented?	0.42 (0.23, 0.72)
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	0.74 (0.45, 0.87)
9. Were the methods used to combine the findings of studies appropriate?	0.45 (0.12, 0.70)
10. Was the likelihood of publication bias assessed?	0.88 (0.75, 0.94)
11. Were potential conflicts of interest included?	0.92 (0.83, 0.96)

Abbreviations: AMSTAR, a measurement tool to assess systematic reviews; CI, confidence interval.

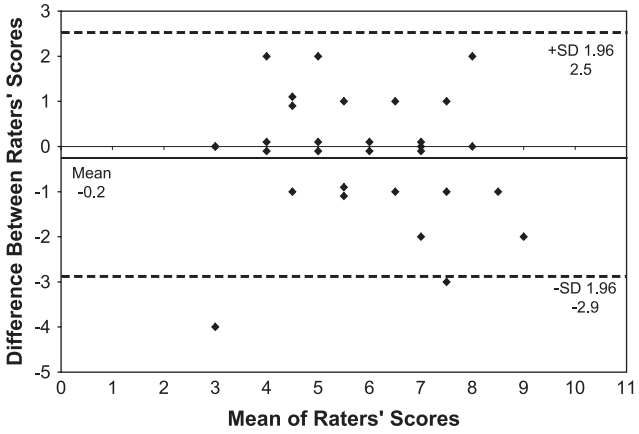

Fig. 1. Bland and Altman plot of interrater agreement on a measurement tool to assess systematic reviews' total score.

The qualitative analysis of the data on agreement led us to make minor modifications to the wording of some items. In particular, under the original item regarding publication bias, the wording was changed to clarify the purpose of the question, that is, to ask whether the status of publication was used as an inclusion criterion (see item 4 and footnote in Appendix). Additional available electronic databases were also added to the question on literature searching (item 3) and I^2 was added to the item on methods used to combine findings (item 9).

3.2. Construct validity

Expressed as a percentage of the maximum score, the results of AMSTAR showed convergence with the results of the other instruments. ICC for AMSTAR was 0.66 (95% CI: 0.28, 0.84) against OQAQ and 0.83 (95% CI: 0.64, 0.92) against Sacks' instrument. The ICC obtained when comparing OQAQ with Sacks' instrument was 0.86 (95% CI: 0.70, 0.93).

3.3. Feasibility

AMSTAR proved easy to apply, each review taking 14.9 (95% CI: 17.0, 12.8) minutes to complete. OQAQ took, on average, 20.3 (95% CI: 22.5, 18.0) minutes to complete, and Sacks' instrument 34.4 (95% CI: 37.3, 31.6) minutes ($P < 0.0001$ for comparison between the three instruments). Two of the reviewers expressed difficulty with

scoring item 4 on publication status: ''was the status of publication (i.e., grey literature) used as an inclusion criterion?''

4. Discussion

There has been a continued proliferation of (largely unvalidated) scales and checklists for assessing the quality of systematic reviews [44]. This causes confusion for those who use reviews in making clinical and policy decisions and who need to be able to distinguish good- from poor-quality reviews. There is a need for a reliable and valid quality assessment instrument that is easy to use. AMSTAR was developed to meet this need. Our aim was not to devise a truly original instrument, but to develop and update the best available from the published literature. Our scan indicated that the OQAQ developed by Oxman and Guyatt [3] and the rating scale of Sacks et al. [4] were among the best out of more than two dozen instruments assessed by us. We found that both instruments had been rigorously developed, but were dated in some respects. We decided to improve the descriptors of the items resulting in the ''enhanced OQAQ'' that we have applied in the subsequent studies [2]. The checklist developed by Sacks et al. showed good quality and was especially comprehensive but unwieldy in general use [4]. We based AMSTAR on a development of both of these instruments. Full details of this development process are published elsewhere [5].

This study suggests that AMSTAR has good agreement, reliability, construct validity, and feasibility to assess the quality of systematic reviews. Its performance in terms of agreement and reliability was similar to OQAQ and better than Sacks' instrument; it adds important items that are not present in either instrument (e.g., publication status, conflict of interest), and has better feasibility than OQAQ or Sacks' instrument. We think AMSTAR can be applied to a wide variety of systematic reviews, but recognize that it has only been tested on systematic reviews of randomized control trials evaluating treatment interventions. We accept that the relatively high reliability of total scores for AMSTAR and OQAQ may be partly because of the raters' familiarity with both instruments, and this reliability needs to be tested more widely in the field.

AMSTAR showed good (convergent) construct validity in comparison with the two existing instruments. A recently published study concluded that the underlying construction of OQAQ is designed for the assessment of meta-analyses. Thus, it is difficult for any other type of review to score highly on the OQAQ, and if the review does not have a meta-analysis component, it may be deemed to have major flaws [45]. AMSTAR can be scored both individually (components) and as a checklist by summing the item scores (overall score). It was psychometrically developed to score each item as if it was not related to the others. Each component came out separate in the factor analysis. Therefore, all reviews have equal chance of scoring well, but meta-analysis will score slightly higher in the overall results.

The feasibility of AMSTAR is documented in terms of the time required to complete an assessment while using it: about 10–15 minutes, which is substantially less than the time needed to complete the other instruments.

In this study, we did not assess the external validity of AMSTAR. This has been carried out separately and the results have recently been published. In that analysis, we looked at differences in overall scores and compared them with the global assessments made by an informed panel. We found very good correlations on the total scores [6].

There has been considerable discussion regarding the merits of using individual component scores or summary scores when assessing systematic reviews and individual studies [45,46]. From a methodological standpoint, it is worth assessing the component scores as they measure different elements, and some may be more important than others in particular situations. Hence, a summary score may obscure important strengths or weaknesses. In the case of AMSTAR, we have tried to develop an instrument that is pragmatic and of value to decision makers. It is valid and easy to use and the total score is meaningful. By this, we mean that we have ensured that the components are nonoverlapping and have separately validated the total score against an external standard [5,6]. Hence, we believe that decision makers can have some confidence that the component scores measure different domains of quality and the overall score is meaningful.

We feel that the main advantage of AMSTAR over the OQAQ and Sacks' instruments lies in its better compromise between comprehensiveness and feasibility. It adds relevant dimensions to those covered in the OQAQ without becoming unwieldy, as with the Sacks' instrument. For example, the item ''sources of support'' was included in the original data set and came out as a component in the factor analysis. Some may doubt the usefulness of the item concerning conflict of interest. We put a lot of thought into the name and description of this item, using all available empirical evidence. In addition to the research previously discussed on this topic, more recent studies also suggest that funding source influences outcomes and quality of research [47]. In a study by Biondi-Zoccai [48], the authors concluded that reviewers who reported previous not-for-profit funding were more likely to carry out higher-quality systematic review. We believe that funding sources are associated with bias in systematic reviews and it is important to rate this aspect of their conduct.

The inclusion of unpublished studies is rated by AMSTAR. This remains a controversial topic in systematic reviews. Several examples in the drugs' field have demonstrated the crucial importance of including nonpublished data. In the case of two Cox-2 inhibitors, rofecoxib and celecoxib, the incomplete reporting of vascular deaths and gastrointestinal events skewed the results of trials that were used as the basis of important decisions [49]. More recently, several systematic

reviews/meta-analyses of antidepressant drugs have shown that the inclusion of unpublished data dramatically alters the perception of benefit in a negative way [50,51]. Based on these examples, we felt it important to include an item in AMSTAR dealing with unpublished studies.

During the development of AMSTAR, careful consideration was given to the wording of the individual items and minor adjustments were made where necessary; despite this process, agreement between observers was disappointingly low on three items. One of these items assessed publication restriction. After discussion between observers, we reworded the descriptor slightly and this has improved agreement. The reworded version is provided in Appendix. The other two items describe ''report of assessment of scientific quality'' and ''appropriate method to combine studies.'' Agreement was low on similar items in the other instruments assessed here. Subjective judgment comes into play when one is asked to assess whether quality of included studies was assessed adequately. Conceivably, one could increase reliability of assessment by providing more detailed instructions or by adding more items or criteria. This would, however, decrease feasibility. It should also be noted that overall agreement on these items was good; hence, their relatively low kappa values are likely caused by skewness in the responses, that is, most of the responses in either the ''yes'' or the ''no'' category. This is a well-known limitation of the kappa statistic [52].

Our study has other limitations. We did not compare AMSTAR with the current state-of-the-art reporting quality of meta-analysis (QUOROM) [53]. The reason for this is that QUOROM is not specifically designed to assess methodological quality. Rather, it is specifically focused on the quality of *reporting* (not conduct) of the review. This does not detract from the utility of QUOROM, but its limited focus made it unsuitable for our study. A further limitation of the present study is the fact that the sample of reviews used is derived from the original source used to develop AMSTAR (Table 2), and one of the assessors is the principal investigator. Thus, application to other reviews and by other assessors is necessary to discover the full potential of this tool. Finally, the number of reviews used to validate AMSTAR was rather small.

Our new instrument builds on previous work. Methodologists continue to struggle with methodological quality issues, whereas decision makers struggle with the challenge of basing policy, clinical, or resource planning decisions on the available evidence. The personal feedback received on AMSTAR has been supportive. AMSTAR is now being used by a number of groups, including the Canadian Agency for

Table 2
Characteristics of included studies

Author	Year	Journal type	Topic area
1. Anonymous	1989	*NEJM*	Breast cancer
2. Appel	1993	*Arch Intern Med*	Blood pressure reduction
3. Buring	1988	*Rev Inf Dis*	Aminoglycoside antibiotics
4. Chalmers	1977	*NEJM*	Acute myocardial infarction
5. Clagett	1988	*Ann Surg*	Venous thromboembolism
6. Counsell	1996	*Cochrane*	Carotoid surgery
7. Daya	1996	*Cochrane*	FSH and HMG in IVF
8. Duley	1996	*Cochrane*	Anticonvulsants for pre-eclampsia
9. Fanning	1992	*Obstet Gynecol*	Ovarian Carcinoma
10. Gent	1986	*Chest*	Deep vein thrombosis
11. Gotzsche	1995	*BMJ*	Esophageal varices
12. Graves	1996	*Cochrane*	Malaria vaccines
13. Henderson	1989	*Ann Intern Med*	Coronary artery bypass surgery
14. Hodnett	1996	*Cochrane*	Delivery settings
15. Hofmeyr	1996	*Cochrane*	Abdominal decompression
16. Hopfenmuller	1994	*Arzneim-Forsch*	Hirnleistungsstorungen im Alter
17. Hughes	1992	*Fertil Steril*	In vitro fertilization and gamete intrafallopian transfer
18. Kaufmann	1988	*Health Psychol*	Hypertension
19. Kramer	1996	*Cochrane*	Lactation
20. Lycka	1990	*Int J Dermatol*	Postherpetic neuralgia
21. McGrath	1996	*Cochrane*	Tardive dyskinesia
22. Mulrow	1988	*JAMA*	Congestive heart failure
23. Ohlsson	1989	*Am J Obstet Gynecol*	Preterm premature rupture of the membranes
24. Perez-Escamilla	1994	*Am J Public Health*	Breastfeeding
25. Renfrew	1996	*Cochrane*	Infant discharge times
26. Renfrew	1996	*Cochrane*	Breastfeeding and early contact
27. Soares	1996	*Cochrane*	Tardive dyskinesia and GABA
28. Thacker	1985	*BJOG*	Imaging ultrasound
29. Velanovich	1989	*Surgery*	Resuscitation
30. Wilson	1992	*J Hosp Infect*	Surgical prophylaxis

FSH, follicle stimulating hormone; hMG, human menopausal gonadotrophin; IVF, in vitro fertilization; GABA, gamma-aminobutyric acid.

Drugs and Technologies in Health, and The Cochrane Effective Practice and Organization of Care Group (EPOC) [54]. With publication of data on reliability and validity in a peer-reviewed journal, we hope that it will help many reviewers with their task of assessing the methodological quality and incorporating the results into their systematic reviews.

In summary, AMSTAR is an empirically developed instrument for documenting the quality of systematic reviews. It was found to have good agreement, reliability, and construct validity in a limited test setting. It combines in one instrument a level of comprehensiveness and feasibility not found in existing instruments. We encourage others to test our new instrument on other samples of systematic reviews. Its ongoing application in the assessment of the quality of systematic reviews will provide further confirmation of its utility.

Further validation is needed to replicate the initial promising results, and this should involve a broader range of assessors and a broader range of reviews to assess whether the reliability and validity are confirmed in diverse circumstances.

Appendix: A measurement tool to assess systematic reviews (AMSTAR)

1. Was an "a priori" design provided?	☐	Yes
The research question and inclusion criteria	☐	No
should be established before the conduct of	☐	Can't answer
the review.	☐	Not applicable
2. Was there duplicate study selection and data	☐	Yes
extraction?	☐	No
There should be at least two independent data	☐	Can't answer
extractors and a consensus procedure for dis-	☐	Not applicable
agreements should be in place.		
3. Was a comprehensive literature search	☐	Yes
performed?	☐	No
At least two electronic sources should be	☐	Can't answer
searched. The report must include years and	☐	Not applicable
databases used (e.g., Central, EMBASE, and MEDLINE). Key words and/or MESH terms must be stated, and where feasible, the search strategy should be provided. All searches should be supplemented by consulting current contents, reviews, textbooks, specialized registers, or experts in the particular field of study, and by reviewing the references in the studies found.		
4. Was the status of publication (i.e., grey liter-	☐	Yes
ature) used as an inclusion criterion?	☐	No
The authors should state that they searched	☐	Can't answer
for reports regardless of their publication	☐	Not applicable
type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.[a]		
5. Was a list of studies (included and excluded)	☐	Yes
provided?	☐	No
A list of included and excluded studies should	☐	Can't answer
be provided.	☐	Not applicable
6. Were the characteristics of the included	☐	Yes
studies provided?	☐	No
In an aggregated form, such as a table, data	☐	Can't answer
from the original studies should be provided	☐	Not applicable
on the participants, interventions, and outcomes. The ranges of characteristics in all the studies analyzed, e.g., age, race, sex, relevant socioeconomic data, disease status, duration, severity, or other diseases should be reported.		
7. Was the scientific quality of the included	☐	Yes
studies assessed and documented?	☐	No
"A priori" methods of assessment should be	☐	Can't answer
provided (e.g., for effectiveness studies if the	☐	Not applicable
author(s) chose to include only randomized, double-blind, placebo-controlled studies, or allocation concealment as inclusion criteria); for other types of studies, alternative items will be relevant.		
8. Was the scientific quality of the included	☐	Yes
studies used appropriately in formulating	☐	No
conclusions?	☐	Can't answer
The results of the methodological rigor and	☐	Not applicable
scientific quality should be considered in the analysis and the conclusions of the review, and explicitly stated in formulating recommendations.		
9. Were the methods used to combine the find-	☐	Yes
ings of studies appropriate?	☐	No
For the pooled results, a test should be done to	☐	Can't answer
ensure the studies were combinable, to assess	☐	Not applicable
their homogeneity (i.e., Chi-squared test for homogeneity, I^2). If heterogeneity exists, a random effects model should be used and/or the clinical appropriateness of combining should be taken into consideration (i.e., is it sensible to combine?).		
10. Was the likelihood of publication bias	☐	Yes
assessed?	☐	No
An assessment of publication bias should in-	☐	Can't answer
clude a combination of graphical aids (e.g.,	☐	Not applicable
funnel plot, other available tests) and/or statistical tests (e.g., Egger regression test).		
11. Was the conflict of interest included?	☐	Yes
Potential sources of support should be clearly	☐	No
acknowledged in both the systematic review	☐	Can't answer
and the included studies.	☐	Not applicable

"Can't answer" is chosen when the item is relevant but not described by the authors; "not applicable" is used when the item is not relevant, such as when a meta-analysis has not been possible or was not attempted by the authors.

[a] The original wording for question #4: *Was the status of publication (i.e., grey literature) used/not used as an exclusion criterion?* The authors should state that they searched for reports regardless of their publication type. The authors should state whether or not they excluded any reports (from the systematic review), based on their publication status, language etc.

References

[1] Moher D, Jadad AR, Nichol G, Penman M, Tugwell P, Walsh S. Assessing the quality of randomized controlled trials: an annotated bibliography of scales and checklists. Control Clin Trials 1995;16:62−73.

[2] Shea B, Dubé C, Moher D. Assessing the quality of reports of systematic reviews: the QUOROM statement compared to other tools. In:

Egger M, Smith GD, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. London: BMJ Books; 2001. p. 122–39.
[3] Oxman AD, Guyatt GH. Validation of an index of the quality of review articles. J Clin Epidemiol 1991;44:1271–8.
[4] Sacks H, Berrier J, Reitman D, Ancona-Berk VA, Chalmers TC. Meta--analyses of randomized controlled trials. N Engl J Med 1987;316:450–5.
[5] Shea B, Grimshaw J, Wells G, Boers M, Andersson N, Hamel C, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Methodol 2007;7:10.
[6] Shea BJ, Bouter LM, Peterson J, Boers M, Andersson N, Ortiz Z, et al. External validation of a measurement tool to assess systematic reviews (AMSTAR). PLoS ONE 2007;2(12):e135010.1371/journal.-pone.0001350. Published online December 26, 2007.
[7] Anonymous. Effects of adjuvant tamoxifen and of cytotoxic therapy on mortality in early breast cancer. An overview of 61 randomized trials among 28,896 women. Early Breast Cancer Trialists' Collaborative Group. NEJM 1989;319:1681–92.
[8] Appel LJ, Miller ER, Seidler AJ, Whelton PK. Does supplementation of diet with "fish oil" reduce blood pressure. Arch Intern Med 1993;153:1429–38.
[9] Buring JE, Evans DA, Mayrent SL, Rosner B, Colton T, Hennekens CH. Randomized trials of aminoglycoside antibiotics: quantitative overview. Rev Inf Dis 1988;10:951–7.
[10] Chalmers TC, Matta RJ, Smith H, Kunzler AM. Evidence favouring the use of anticoagulants in the hospital phase of acute myocardial infarction. NEJM 1977;297:1091–6.
[11] Clagett GP, Reisch JS. Prevention of venous thromboembolism in general surgical patients. Results of meta-analysis. Ann Surg 1988;208:227–40.
[12] Counsell C, Warlow C, Naylor R. Different patches in carotoid surgery. Cochrane Library 1996 (issue 3).
[13] Daya S. Comparison of FSH and HMG in IVF. Cochrane Library 1996;(3).
[14] Duley L, Gulmezoglu AM, Henderson-Smart DJ. Anticonvulsants for pre-eclampsia. Cochrane Library 1996;(3).
[15] Fanning J, Bennett TZ, Hilgers RD. Meta-analysis of cisplatin, doxorubicin, and cyclophosphamide versus cisplatin and cyclophosphamide chemotherapy of ovarian carcinoma. Obstet Gynecol 1992;80:954–60.
[16] Gent M, Roberts RS. A meta-analysis of the studies of dihydroergotamine plus heparin in the prophylaxis of deep vein thrombosis. Chest 1986;89:396S–400S.
[17] Gotzsche PC, Gjorup I, Bonnen H, Brahe NE, Becker U, Burcharth F. Somatostatin vs placebo in bleeding oesophageal varices: randomised trial and meta-analysis. BMJ 1995;310:1495–8.
[18] Graves P. Malaria vaccines. Cochrane Library 1996;(Issue 3).
[19] Henderson WG, Goldman S, Copeland J, Moritz TE, Harker LA. Antiplatelet or anticoagulant therapy, after coronary artery bypass surgery: a meta-analysis of clinical trials. Ann Intern Med 1989;111:743–50.
[20] Hodnett ED. Alternative versus conventional delivery settings. Cochrane Library 1996;(3).
[21] Hofmeyr GJ. Abdominal decompression. Cochrane Library 1996;(3).
[22] Hopfenmuller W. Nackweis der therapeutischen Wirksamkeit eines Ginkgo biloba-Spezial extrakes: Meta-Analyse von 11 klinischen Studien bei Patienten mit Hirnleistungsstorungen im Alter. Arzneimittel-Forschung 1994;44:1005–13.
[23] Hughes E, Fedorkow DM, Daya S, Sagle MA, van de Kopple P, Collins JA. The routine use of gonadotropin-releasing hormone agonists prior to in vitro fertilization and gamete intra-fallopian transfer: a meta--analysis of randomized controlled trials. Fertil Steril 1992;58:888–96.
[24] Kaufmann PG, Jacob RG, Ewart CK, Chesney MA, Muenz LR, Doub N, et al. Hypertension Intervention Pooling Project. Health Psychol 1988;7(Suppl):209–24.
[25] Kramer MS. Maternal antigen avoidance as lactation. Cochrane Library 1996;(3).
[26] Lycka BA. Postherpetic neuralgia and systemic corticosteroid therapy. Efficacy and safety. Int J Dermatol 1990;29:523–7.
[27] McGrath JJ, Soares KVS. Tardive dyskinesia and benzodiazepines. Cochrane Library 1996;(3).
[28] Mulrow CD, Mulrow JP, Linn WD, Aguilar C, Ramirez C. Relative efficacy of vasodilator therapy in chronic congestive heart failure. Implications of randomized trials. JAMA 1988;259:3422–6.
[29] Ohlsson A. Treatments of preterm premature rupture of the membranes: a meta-analysis. Am J Obstet Gynecol 1989;160: 890–906.
[30] Perez-Escamilla R, Pollitt E, Lonnerdal B, Dewey KG. Infant feeding policies in maternity wards and their effect on breast-feeding success: an analytical overview. Am J Public Health 1994;84:89–97.
[31] Renfrew MJ, Lang S. Breastfeeding and discharge times. Cochrane Library 1996;(3).
[32] Renfrew MJ, Lang S. Breastfeeding and early contact. Cochrane Library 1996;(3).
[33] Soares KVS, McGrath JJ, Deeks JJ. Tardive dyskinesia and GABA agonist drugs. Cochrane Library 1996;(3).
[34] Thacker SB. Quality of controlled clinical trials. The case of imaging ultrasound in obstetrics: a review. BJOG 1985;92:437–44.
[35] Velanovich V. Crystalloid versus colloid fluid resuscitation: a meta-analysis of mortality. Surgery 1989;105:65–71.
[36] Wilson APR, Shrimpton S, Jaderberg M. A meta-analysis of the use of amoxycillin-clavulanic acid in surgical prophylaxis. J Hosp Infect 1992;22:9–21.
[37] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37–46.
[38] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;i: 307–10.http://www-users.york.ac.uk/~mb55/meas/ba.htm.
[39] Bland JM, Altman DG. Statistical methods for assessing agreement between measurements. Biochim Clin 1987;11:399–404.
[40] Anonymous. This week's citation classic: comparing methods of clinical measurement. Curr Contents 1992; CC/NUMBER 40: 8. Available at http://garfield.library.upenn.edu/classics1992/A1992JN24800001.pdf.
[41] Uebersax JS. Diversity of decision-making models and the measurement of inter-rater agreement. Psychol Bull 1987;101:140–6.
[42] Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol Bull 1968;70: 213–20.
[43] Tugwell P, Bombardier C. A methodological framework for developing and selecting endpoints in clinical trials. J Rheumatol 1982;9:758–62.
[44] Singh S, Bai A, Lal A, Yu C, Ahmed F. Developing evidence-based best practices for the prescribing and use of proton pump inhibitors in Canada. Ottawa, Canada: The Canadian Agency for Drugs and Technologies in Health (CADTH); 2006.
[45] Balk E, Bonis P, Moskowitz H, Schmid C, Ioannidis J, Wang C, et al. Correlation of quality measures with estimates of treatment effect in meta-analyses of randomized controlled trials. JAMA 2002;287:2973–82.
[46] Jüni P, Altman DG, Egger M. Assessing the quality of controlled clinical trials. In: Egger M, Davey Smith G, Altman DG, editors. Systematic reviews in health care: meta-analysis in context. 2nd ed. London: BMJ Books; 2001.
[47] Barnes DE, Bero LA. Why review articles on the health effects of passive smoking reach different conclusions. JAMA 1998;279:1566–70.
[48] Biondi-Zoccai G, Lotrionte M, Abbate A, Testa L. Compliance with QUOROM and quality of reporting of overlapping meta-analyses on the role of acetylcysteine in the prevention of contrast associated nephropathy: case study. BMJ 2006;332:202–6.
[49] Chou R, Helfand M. Challenges in systematic reviews that assess treatment harms. Ann Intern Med 2005;142:1090–9.
[50] Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. Selective Publication of antidepressant trials and its influence on apparent efficacy. N Eng J Med 2008;358:252–60.
[51] Whittington CJ, Kendall T, Fonagy P, Cottrell D, Cotgrove A, Boddington E. Selective serotonin reuptake inhibitors in childhood depression: systematic review of published versus unpublished data. Lancet 2004;363:1341–5.
[52] McGinn T, Guyatt G, Cook R, Meade M. Diagnosis: measuring agreement beyond chance. In: Guyatt G, Rennie D, editors. Users'

guide to the medical literature. A manual for evidence-based clinical practice. Chicago, IL: AMA Press; 2002. p. 461−70.

[53] Moher D, Cook DJ, Eastwood S, Olkin I, Rennie D. Stroup D for the QUOROM group. Improving the reporting quality of meta-analysis of randomized controlled trials: the QUOROM statement. Lancet 1999;354:1896−900.

[54] Oxman AD, Schünemann HJ, Fretheim A. Improving the use of research evidence in guideline development: synthesis and presentation of evidence. Received April 7, 2006, Accepted December 5, 2006. Available at. Health Res Policy Syst 2006;(4): 2010.1186/1478-4505-4-20. http://www.health-policy-systems.com/content/4/1/20.